Philos Stud (2014) 167:273–290
DOI 10.1007/s11098-012-0087-1

Indirect representation and the self-representational theory of consciousness

Ben Phillips

Published online: 8 January 2013

Abstract According to Uriah Kriegel's self-representational theory of consciousness, mental state M is conscious just in case it is a complex with suitably integrated proper parts, M_1 and M_2, such that M_1 is a higher-order representation of lower-order representation M_2. Kriegel claims that M thereby "indirectly" represents itself, and he attempts to motivate this claim by appealing to what he regards as intuitive cases of indirect perceptual and pictorial representation. For example, Kriegel claims that it's natural to say that in directly perceiving the front surface of an apple one thereby perceives the apple itself. Cases such as this are supposed to provide intuitive support for the principle that if X represents Y, and Y is highly integrated into complex object Z, then X indirectly represents Z. In this paper I provide counterexamples to Kriegel's principle of indirect representation, before going on to argue that we can explain what is going on in those cases in which the subject seems to represent a complex whole by representing one its parts without positing *indirect* representations anyway. I then argue that my alternative approach is superior to Kriegel's in a number of ways, thereby rendering his theory of consciousness implausible.

Keywords Consciousness · The self-representational theory of consciousness · The HOT theory of consciousness · Indirect representation · Perceptual representation · Complexes

1 Introduction

According to the theory of consciousness Uriah Kriegel (2006, 2007, 2009a, b, 2012a, b) has been developing recently, conscious mental states "indirectly"

B. Phillips (✉)
The Graduate Center, CUNY and The Saul Kripke Center, 365 Fifth Avenue, New York, NY 10016-4309, USA
e-mail: ben.s.phillips@gmail.com

represent themselves. More specifically, on Kriegel's view, mental state M is conscious just in case it is a complex with suitably integrated proper parts, M_1 and M_2, such that M_1 is a higher-order representation of lower-order representation M_2 (e.g. a perceptual representation). Kriegel attempts to motivate the claim that conscious state M thereby *indirectly* represents itself by appealing to what he regards as intuitive cases of indirect perceptual and pictorial representation. For example, Kriegel claims that it is natural to say that in directly perceiving the front surface of an apple one thereby perceives the apple itself. Cases such as this are supposed to provide independent, intuitive, support for the principle that if X represents Y, and Y is highly integrated into complex object Z, then X indirectly represents Z—call this "the principle of indirect representation" (hereafter "(IR)").

In this paper I raise several problems with Kriegel's appeal to the notion of indirect representation. In Sect. 3 I argue that there are cases wherein (IR) doesn't seem to hold; more specifically, cases wherein the subject seems to be completely unaware of the complex object that she is supposed to be indirectly representing according to (IR). Then, in Sect. 4, I consider potential ways to amend (IR), arguing that each faces its own counterexamples.

In Sect. 5 I provide an alternative to Kriegel's approach that avoids any commitment to indirect representations. In particular, I argue that in those cases in which the subject seems to represent the complex whole by directly perceiving one of its parts, her direct perception of the part *causes* her to token a distinct, direct, representation of the complex whole. I extend this approach to the case of thoughts before outlining an approach to the pictorial representation of complex wholes that also avoids any commitment to indirect representations.

Finally, in Sect. 6, I argue that my approach to the representation of complex wholes is superior to Kriegel's for it explains why it is that in those kinds of cases given in Sects. 3 and 4, where the subject seems to be completely unaware of the relevant complex's existence, intuitive support for the notion of indirect representation is lacking. Furthermore, I argue that the notion of indirect representation has no power when it comes to explaining behavior—a problem that my approach avoids.

The moral will thus be that Kriegel's theory of consciousness is problematic because it requires us to commit to an implausible view of how objects and the complex wholes into which they are integrated get represented by perceptions, pictures and thoughts.

2 HOTs and self-representing mental states

Before outlining Kriegel's self-representational theory of consciousness, it will be instructive to very briefly review the HOT theory of consciousness—the self-representational theory's closest cousin. Briefly put, the HOT theory says that my mental state, M, is conscious just in case I have an appropriate higher-order thought that I am in M.[1] What counts as an "appropriate" higher-order thought? According to Rosenthal, the consciousness-conferring thought must be roughly contemporaneous

[1] See Rosenthal (1986, p. 335; 1990, pp. 36–37; 1993, p. 199; 2002, p. 410; and 2005, p. 26).

with *M*, but it mustn't be mediated by any conscious inference or observation (2002, pp. 408–409; 2005, pp. 183–184).[2,3]

In order to locate the point at which the self-representational theory diverges from the HOT theory, notice that according to the HOT theorist, mental state *M* is conscious just in case it is targeted by an appropriate HOT, *T*, *where M and T are distinct mental states*. In contrast, the self-representationalist claims that rather than being represented by some state other than *M*, *M* "represents itself" in some sense.

2.1 Naturalizing self-representation

How is *M* supposed to achieve this feat of self-representation? If we take the claim that it "represents itself" at face value then we get the view according to which my mental state, *M*, is conscious just in case I'm in an appropriate mental state, *M**, such that *M** represents me as being in *M*, and $M = M^*$. Call this brand of self-representation "pure self-representation".

Kriegel (2009a, pp. 205–215) doubts that (mental) pure self-representation can be naturalized.[4] In brief, the worry is that the main naturalistic accounts of mental representation on the market attempt to ground it in natural relations holding between mental states and certain other entities (e.g. causal ones), where these relations are not of the kind that can hold between a state and itself.[5] I won't attempt to determine the prospects for naturalizing pure self-representation here. Rather, I'll follow Kriegel and assume that the prospects are dim, and that the self-representationalist should therefore find another way to accommodate self-representing mental states.

If pure self-representation is not naturalistically respectable then where does this leave the self-representationalist? According to Kriegel, there is a way to retain the view that conscious state *M* self-represents in a naturalistically respectable fashion. Kriegel's view is as follows:

> *S*'s mental state, *M*, is conscious iff *M* is a complex with proper parts, M_1 and M_2, such that (a) M_1 is an appropriate (occurrent) representation of lower-order representation M_2 (M_2 is either a world-directed representation or a higher-order representation itself) and (b) M_1 thereby represents *M* (Kriegel 2006, p. 153; 2009a, pp. 222–223).[6]

[2] Rosenthal also claims that the consciousness-conferring thought (i) needn't be conscious itself (2002, p. 410), and (ii) needn't accurately represent the target state (2002, p. 415); in fact, according to Rosenthal (2011), the target state needn't exist.

[3] For some alternative higher-order theories of consciousness see Armstrong (1968, 1984), Carruthers (1996, 2000, 2005) and Lycan (1996).

[4] Perhaps sentences of natural language can purely self-represent; e.g. consider 'This sentence has thirty-one letters'. The problematic claim that concerns us here is the claim that *mental* pure self-representation can be naturalized.

[5] Even if the prospects for naturalizing (mental) pure self-representation are bad, it doesn't seem *intuitively* obvious that mental states cannot self-represent in the pure sense. For instance, it's not intuitively obvious to me that I can't have a purely self-representing thought of the form *THIS VERY THOUGHT IS OCCURRING NOW*.

[6] Gennaro (2002, 2004, 2005, 2006) defends a theory similar to Kriegel's except for the fact that he denies that self-representing conscious states exhibit *indirect* representation (i.e. he rejects (b)). I will leave aside the question of whether Gennaro should be seen as a self-representationalist.

Two obvious questions need addressing here: First, what is a *complex* mental state? Second, how is it that in representing M_2, M_1 *thereby* represents M?

According to Kriegel, a "complex" is a sum whose parts are essentially related in some way. That is, whereas the existence of parts A and B is sufficient for the *sum* of A and B to exist, A and B have to stand in a certain relation for the *complex* of A and B to exist (2009a, pp. 221–222). For example, a human being is a complex of molecules, rather than a mere sum of molecules, because the molecules must be arranged in a certain way in order for the human being to exist. In the same way, under the view endorsed by Kriegel (2006, pp. 150–151; 2007; 2009a, pp. 221–223), conscious state M is a *complex* of M_1 and M_2. According to Kriegel, what makes M_1 and M_2 part of a single complex mental state is "the fact that they are integrated and unified through a psychologically real cognitive process of information integration" (2006, p. 150). He calls this kind of integration "cross-order integration" (2009a, p. 236).[7]

How exactly does M_1 represent M *in virtue of representing* M_2? Kriegel answers this question by appealing to the notion of indirect representation:

> …we may say that a painting depicts a house even though a portion of the house is occluded by a bush in the corner. It is natural to describe this as a case where a painting represents an entire house by, or in virtue of, representing a big part of it; the entire house is represented indirectly, its big part is represented directly. Likewise, a visual perception may represent an apple by representing its front surface and an olfactory perception may represent an apple pie by representing its odor; the surface and odor are directly represented, the apple and pie indirectly represented (2009a, p. 225).

According to Kriegel, the "in virtue of" locution behaves the same way in the representational context as it does in other contexts. For instance, it seems natural to say that I dented the car *in virtue of* denting the door, and that I'm in my house *in virtue of* being in my living room (2009a, p. 225). Of course, we wouldn't want to say that in all cases, if X represents Y and Y is part of Z then X represents Z, for there are clear counterexamples to this principle. For instance, my fingernail is part of the Earth but a representation of my fingernail does not count as a representation of the Earth.

Thus in order for X to represent Z in virtue of representing part of Z, some further condition must be met. Kriegel's suggestion is as follows:

[7] Kriegel speculates that the mechanism realizing cross-order integration is similar to the mechanism that realizes feature binding in perception. More specifically, noting the popularity of the neural synchrony model of feature binding in perception, Kriegel speculates that cross-order integration arises due the synchronization of the firing rates of those neurons corresponding to the higher-order representation with those corresponding to the relevant lower-order representation (Kriegel 2007, 2009a, pp. 243–248). This is not the place to determine whether Kriegel's hypothesis about the mechanism realizing cross-order integration is plausible or not, but notice that an immediate problem arises. In the case of feature binding in perception, the kind of unity that is supposed to arise in virtue of synchronous firing is the representation of various distinct features as being instantiated by a single object, whereas in the case of cross-order integration, the kind of unity in question involves the fact that the relevant higher- and lower-order representations belong to the same mental state. How could the same kind of mechanism give rise to one kind of unity in the case of feature binding and a different kind of unity in the cross-order case?

(IR) *X* represents *Z* in virtue of representing *Y* if (i) *X* represents *Y*, (ii) *Y* is part of *Z*, and (iii) *Y* is highly integrated into *Z* (2009a, p. 227).

(IR) captures the intuition that if I perceive the front surface of an apple then I perceive the apple itself, for the front surface of the apple is highly integrated into the apple, whereas in perceiving my fingernail my mental state does not represent the Earth for my fingernail is not highly integrated into the Earth—or so the argument goes. Kriegel admits that the notion of "high integration" is somewhat "vague and obscure"; however he registers his confidence that "something *like* the above analysis is probably right and more precisely expressible" (2009a, p. 227).

Armed with the notion of indirect representation, Kriegel claims that we are now in a position to understand how it is that conscious mental state *M* represents itself in a naturalistically respectable fashion. For if, as Kriegel claims, *M* is a complex, consisting of highly integrated parts, M_1 and M_2, where M_1 *directly* represents M_2, all we have to do is apply our favored naturalistic semantics to the *direct* representation of M_2 by M_1 and, given (IR), we have thereby secured a naturalistic account of *M*'s indirect self-representation.[8]

3 Counterexamples to the principle of indirect representation

Does (IR)—the crucial principle underpinning Kriegel's theory of consciousness—really enjoy independent, intuitive, support? Consider the perceptual case. Is there really a robust intuition that perceptual states obey (IR)? For instance, suppose a child, who has never acquired the concept of a cell, looks through a microscope and sees only part of a cell. Suppose that, due to the level of magnification, she is only receiving sensory input from the cell's nucleus, which has been stained blue. Given that the child is completely unaware that the object she perceives is part of a complex cell, is there really an intuition that her perceptual representation of the blue splotch also represents the cell of which it happens to be a part? It seems not. The inclination is to say that the child is not in a state that represents the cell at all.[9]

Now consider a case involving a kind of object that isn't usually integrated into the relevant kind of complex. For instance, suppose that, unbeknownst to you, your housemate has taken your lemon from the fruit bowl and has used it to construct a lemon battery (a cell battery consisting of a lemon with two electrodes attached to it). The lemon is a large part of the battery and realizes a crucial function (it constitutes the battery's ion bridge), thereby making it highly integrated into the lemon battery. Now suppose you walk past the kitchen and see the lemon but you don't see the electrodes that are attached to it—they're out of your line of sight and

[8] Of course, there are arguments for and against self-representationalism that I will not consider here. See Kriegel (2006, 2007, 2009a, b, 2012a, b) for arguments in favor of self-representationalism. See Weisberg (2008) for an argument that one of the key advantages that the self-representational theory is often seen as having over the HOT theory is illusory. See Gertler (2012), Brogaard (2012), Van Gulick (2012), and Kriegel (2012a) for a discussion of arguments for and against Kriegel's specific version of the self-representational theory.

[9] I'm setting aside the issue of whether one can perceive things through a microscope—this is irrelevant for my purposes here.

so none of the light waves reflected from them impinge on your retina. Do you thereby represent the entire complex (i.e. the battery) into which the lemon is highly integrated? It seems highly counterintuitive to say that you do. All you seem to represent is the lemon.

What goes for the perceptual case goes for the pictorial case too. For instance, suppose you go into your bedroom and draw a picture of the lemon after having just directly perceived it. When it comes to the claim that your picture represents the battery, not just the lemon, intuitions seem to go against (IR) here too.

Now consider the case of thoughts.[10] Is there intuitive support for the claim that thoughts obey (IR)? Suppose you're a particle physicist who has just detected an electron in your laboratory, thereby giving rise to the thought *THAT ELECTRON HAD AN UNEXPECTED VELOCITY*. Moreover, suppose that, unbeknownst to you, this electron was highly integrated into a molecule that went undetected by your measuring apparatus. Moreover, suppose you weren't expecting to detect a molecule of any kind. Were you in a mental state that represents the molecule into which the electron was integrated? I have a strong inclination to say that your mental state did not represent the molecule at all.

4 Can (IR) be fixed?

The examples I just gave are designed to show that intuitions do not in fact support the view that in perceiving, pictorially representing or thinking about object *O*, where *O* happens to be highly integrated into some complex whole, the whole itself must be represented as well. This therefore casts serious doubt on Kriegel's claim that there is independent, intuitive, support for (IR)—the crucial principle underpinning his theory of consciousness.

Can we amend (IR) though? Perhaps the counterexamples I provided above show that (IR) merely provides necessary conditions for indirect representation, and that we can uncover the missing sufficient conditions on Kriegel's behalf. In what follows I'll argue that the prospects for augmenting (IR) in this fashion are not good. I'll then go on to provide an alternative account of what is going on in those cases in which the subject seems to represent the whole *by* directly representing the part, where this account eschews the notion of indirect representation altogether.

4.1 Bracketing the subject's expectations

How might we amend (IR) in order to accommodate the conviction that in the cases I provided above, the subject is not in a state that represents the relevant complex? The first thing to notice about these counterexamples is that they all involve subjects who, in some sense yet to be specified, don't *expect* the relevant complex to be present. Furthermore, this lack of expectation on the subject's part is what seems to

[10] Kriegel doesn't attempt to give any intuitive cases of *thoughts* exhibiting indirect representation. Rather, he attempts to motivate (IR) via cases involving perceptual and pictorial representation alone; the thought presumably being that this will be enough to motivate (IR) for every kind of intentional state.

be driving the intuition that she is not in a state that represents the given complex. For example, given that you weren't aware that your housemate had been using your lemons to construct lemon batteries, you didn't expect the lemon that you perceived to be part of a lemon battery, and this is why the intuition is that you weren't in a state that represents the lemon battery when you directly perceived the lemon. Similarly, in the nucleus/cell case, the child didn't expect the blue splotch to be part of a cell, while in the electron/molecule case, you weren't expecting the electron you detected to be part of a molecule.

Given that our intuitions about whether the complex in question is represented by the subject seem to be sensitive to facts about her expectations, any attempt to bracket these facts in giving sufficient conditions for indirect representation will face counterexamples. For instance, suppose we were to try augmenting (IR) by adding condition (iv) as follows:

> (IR_U) *X* represents *Z* in virtue of representing *Y* if (i) *X* represents *Y*, (ii) *Y* is part of *Z*, (iii) *Y* is highly integrated into *Z*, and (iv) objects of the same kind as *Y* are *usually* highly integrated into objects of the same kind as *Z*.

Something along the lines of (IR_U) may well accommodate our intuitions about some of the cases we've considered so far. For example, given that apple surfaces are *usually* highly integrated into apples, (IR_U) accommodates the intuition that in representing the apple's front surface you represent the apple. And given that lemons are *not* usually highly integrated into lemon batteries, (IR_U) accommodates the intuition that in the case described above, you were not in a mental state that represents your housemate's lemon battery.[11]

However, (IR_U) does not accommodate our intuitions about the other cases described above. This is because (IR_U) brackets facts about the subject's expectations concerning the presence of the relevant complex. Take the case in which the child directly perceives the blue splotch (i.e. the nucleus) but fails to realize that it's part of a cell. Even if we specify that nuclei are *usually* highly integrated into cells, the intuition remains that the child is not in a state that represents the cell. Similarly, even if we specify that electrons are *usually* highly integrated into molecules in the relevant kinds of circumstances, the intuition that you were not in a state that occurrently represents the molecule remains.

The lesson is that in trying to amend (IR) we can't afford to bracket all facts concerning the subject's expectations. For if we propose some sufficient conditions for indirect representation, where these conditions don't have anything at all to say about the subject's expectations concerning the presence of the relevant complex, then there will be counterexamples to the proposed principle in which the subject doesn't seem to represent the complex because she doesn't expect it to be present when she perceives the relevant part.

[11] Of course, we would need to say what 'usually' means. I'll set this issue aside though for my concern here is with the general problem faced by any approach that brackets facts concerning the subject's expectations.

4.2 Taking the subject's expectations into account

How might we amend (IR) in such a way that facts concerning the subject's expectations are appropriately incorporated? There are two obvious strategies available. One strategy is to modify (IR_U) by specifying that the subject is somehow aware of the fact that objects of the relevant kind are usually highly integrated into complexes of the relevant kind.[12] This gives us the following variant of (IR_U):

> (IR_U*) *S*'s state, *X*, occurrently represents *Z* in virtue of occurrently representing *Y* if (i) *X* occurrently represents *Y*, (ii) *Y* is part of *Z*, (iii) *Y* is highly integrated into *Z*, and (iv) *S believes that* objects of the same kind as *Y* are usually highly integrated into objects of the same kind as *Z*.[13]

Another strategy for capturing the sense in which the subject 'expects' the relevant kind of complex to be present is to simply appeal to her *disposition* to directly represent the complex as a result of directly representing the relevant part. This gives us the following version of the principle of indirect representation:

> (IR_D) *S*'s state, *X*, occurrently represents *Z* in virtue of occurrently representing *Y* if (i) *X* occurrently represents *Y*, (ii) *Y* is part of *Z*, (iii) *Y* is highly integrated into *Z*, and (iv) *S* is disposed to represent *Z* as a result of representing *Y*.

On this approach, *S*'s background belief that objects of the same kind as *Y* are usually highly integrated into objects of the same kind as *Z* may well be (partly) responsible for her disposition to represent *Z* as a result of representing *Y*. However, this background belief is not construed as constitutive of *S*'s indirect representation of *Z*.[14]

Do (IR_U*) and (IR_D) adequately accommodate our intuitions about the cases that proved problematic for Kriegel's original version of the principle? Take the case of the lemon battery. (IR_U*) and (IR_D) both accommodate the intuition that you didn't represent your housemate's lemon battery upon directly perceiving the lemon for, at the time, you didn't have the requisite background belief regarding lemon batteries and thus you weren't disposed to represent a lemon battery upon directly perceiving a lemon. Similarly, the absence of any requisite background beliefs or dispositions concerning complexes of the relevant kind means that (IR_U*) and (IR_D) both

[12] Thanks to an anonymous referee for suggesting this strategy.

[13] In some cases *S*'s belief—as specified in (iv)—might count as merely *implicit* or *dispositional*. In other cases, *S*'s belief may well count as *explicit* and *occurrent*. This is in contrast to *X*, which must be an occurrent (thought-like) representation on Kriegel's view—I've made this explicit in the formulation of (IR_U*) above. I'll say more about the nature of *S*'s belief below in Sect. 4.3.

[14] There are variations of (IR_U*) and (IR_D) that I won't explicitly discuss here. For instance, once condition (iv) of (IR_U*) is included in our principle of indirect representation, the question as to whether (ii) and (iii) are necessary arises. For example, you know that façades are usually highly integrated into houses, but what do we say about the rare situation in which you're perceiving a façade that, unbeknownst to you, is not attached to a house? There's a temptation to say that you indirectly represent the complex in this kind of case (i.e. a house) even though no such complex exists, contrary to what (ii) and (iii) presuppose. The same point applies to (IR_D). I won't explicitly discuss variations of (IR_U*) and (IR_D) such as this, for they all face the same counterexamples that I give below.

accommodate the intuition that the child did not represent the cell upon directly perceiving the blue splotch, as well as the intuition that you did not represent the molecule upon directly representing the electron in thought.[15]

4.3 Counterexamples to (IR_U*) and (IR_D)

So (IR_U*) and (IR_D) succeed in accommodating our intuitions about what gets represented in those cases that constituted counterexamples to Kriegel's original version of the principle of indirect representation. But (IR_U*) and (IR_D) face their own counterexamples. To see why, suppose we modify each of the cases given above by stipulating that the subject does in fact have the requisite background belief (as specified in (IR_U*)), which means that she is thereby disposed to directly represent the complex in question (as specified in (IR_D)). Does making such a stipulation thereby give rise to the conviction that the subject occurrently represents the given complex in each case?

For instance, take the case in which you detect the electron and token the thought *THAT ELECTRON HAD AN UNEXPECTED VELOCITY*. If we stipulate that, at the time, you had a disposition to token a direct representation of the molecule (because of your background belief that electrons of the relevant kind are usually highly integrated into molecules of the relevant kind), does this give rise to the intuition that you were in a state that occurrently represents the molecule?

In order to answer this question we first need to specify whether or not your background belief caused you to token a direct representation of the complex whole or not? Suppose it did. Then, of course, we needn't posit any *indirect* representation in order to accommodate the intuition that you were in a state that represents the molecule, for the *direct* representation that you tokened will do the job. Now suppose that your disposition to token a direct representation of the molecule was *not* manifested for some reason. For instance, suppose a loud noise distracted you to such a degree that your background belief about the relevant kind of molecule lay dormant and failed to cause you to token a direct representation of the molecule. What do our intuitions say about this kind of case? Were you in a state that occurrently represents the complex whole in addition to the part?

As with the original case in which you weren't even disposed to directly represent the molecule upon thinking about the electron, the inclination is to say that in the case in which you have the relevant disposition *but fail to manifest it*, you're not in a state that occurrently represents the molecule. If it is left unspecified whether your disposition was manifested or not, then we may well be tempted to think that you probably occurrently represented the molecule due to the fact that you probably manifested your disposition. But this inclination is removed once it is

[15] It's not clear how (IR_U*) and (IR_D) could be made to work in the case of pictorial representation, for whom would we identify as the subject appealed to in (iv)? Could we specify that the subject is the *viewer* of the picture*?* Alternatively, could we say that the subject is the picture's *creator*? Answering these questions would take us too far afield. In any case, even if there's a plausible way to construe pictorial representations as obeying (IR_U*) or (IR_D), I argue below that we don't have to look beyond thoughts and perceptions to find counterexamples to any principle akin to these variants of (IR).

specified that your disposition was not manifested because your background belief (as specified in (IR_U*)) lay dormant for some reason.

It seems safe to say that from the first-person perspective you wouldn't take yourself to be in a state that occurrently represents the molecule, for it was stipulated that your background belief about the relevant kind of molecule lay dormant and thus your disposition to token a direct representation of the molecule-complex was not manifested. Moreover, I'll argue below in Sect. 6 that attributing an occurrent representation of the complex whole to the subject in a case of this kind will not help us to explain any of her behavior. This might also partly explain why we're not inclined to attribute an occurrent representation of the molecule to you in the present case, i.e. because it won't help us to explain any of your behavior from the third-person perspective.

Finally, even if, contrary to what I've claimed, there is an intuition that you possess *some* representation of the molecule (despite the fact that you failed to token an occurrent, direct, representation of it), this won't necessarily help Kriegel. To see why, call your thought about the electron, *T*, and call the state that disposes you to token a direct representation of the molecule, *D*. Under Kriegel's original approach—as well as the variants we're considering—the electron and the molecule are both occurrently represented by *T*. So the question is, would an intuition to the effect that you possess *some* representation of the molecule motivate the claim that *T* is the (occurrent) representation in question? I don't see how it could. It may well be that there's an intuitive sense in which *D implicitly* (or *tacitly*) represents the molecule due its dispositional properties, but this is not what Kriegel needs for his view to work. He needs it to be the case that the molecule is both (i) occurrently represented, and (ii) represented by *T*, not *D*.[16]

It's hard to see what other independent motivation there could be for construing *T* as occurrently representing the molecule. Could it be that *D* causally influences the content of *T* in such a way that *T* comes to occurrently represent the molecule? Of course, the problem with this suggestion is that representations are presumably individuated by their contents, and thus if *D* and *T* were to interact in this way then not only would the resulting representation be a *direct* one, it would not be *T*. In fact, this alternative view of how complex wholes get represented in these kinds of cases is the one I'll defend below.

5 Alternatives to indirect representation

In attempting to amend the principle of indirect representation we've seen that even if the subject is disposed to directly represent complex *C* upon directly representing one of its parts, unless this disposition is manifested we're not inclined to think that she's in a state that occurrently represents *C*. This suggests that the correct account of what's going on in the kinds of cases we've been considering does not require us

[16] Notice that if Kriegel were to accept that the molecule is non-occurrently represented by *D*, as opposed to being occurrently represented by *T*, then this would make his theory of consciousness collapse into a version of the dispositional HOT theory defended by Carruthers (1996, 2000, 2005).

to invoke indirect representations at all. Rather, the considerations above suggest that whether the subject represents the complex in question is determined by whether or not she tokens a *direct* representation of it in addition to her direct representation of the part. In what follows I'll argue for just such an approach. Moreover, I'll argue that apart from its intuitive appeal, this alternative to Kriegel's approach also has the upper hand when it comes to explaining behavior.

5.1 Perception and the two-vehicles approach

Take Kriegel's example in which, according to him, it seems natural to say that in perceiving the front surface of an apple you perceive the apple itself. According to Kriegel, a single vehicle directly represents the front surface of the apple and thereby indirectly represents the apple itself. Alternatively, the approach suggested above says that *distinct* vehicles separately represent the front surface of the apple and the apple itself—call this latter view the "two-vehicles" approach.

In adopting the two-vehicles approach we can remain neutral on the question of where perception ends and thought begins, for the two-vehicles approach that I'm defending here leaves it open whether the two vehicles in question both belong to the same perceptual state or whether one of them belongs to perception, and the other to thought.[17] Whichever way one chooses to go, the commitment to indirect representation is avoided.[18]

Suppose I'm right that in those cases in which the subject seems to represent the complex whole in addition to her direct perception of the part, this is because her direct representation of the part is accompanied by a direct representation of the complex whole. What gives rise to this direct representation of the complex whole?

As has already been pointed out, one obvious explanation for why such a disposition would be possessed by the subject is that it is engendered by her background beliefs. More specifically, in many instances, the accompanying (direct) representation of the complex whole will have been tokened via an unconscious inference. For instance, suppose you're looking at a façade. Given your set of background beliefs—e.g. your belief that most façades are attached to houses; your belief that you're not on a film set; your memory of going into the house to which this particular façade is attached, etc.—the sensory input that gives rise to the representation of a façade causes you to token the accompanying representation of an entire house.

[17] Levine (2010) very briefly suggests the view according to which the part is represented in perception and the whole in *thought*, however he doesn't go on to consider the alternative view according to which part and whole are both represented in perception by distinct vehicles.

[18] Spelke (1988) and Burge (2010, pp. 438–449) are influential proponents of opposing sides in the debate concerning the location of the border between perception and thought. Moreover, it's worth pointing out that despite their disagreement on this matter, they both endorse the view that surfaces and the physical bodies/objects of which they are parts are represented separately in the brain. That is, they both seem to endorse the two-vehicles approach when it comes to surface/object representation. In particular, see Spelke (1988, pp. 229–230) and Burge (2010, pp. 448–449). Marr (1982) is another notable defender of the view that surfaces and the physical objects of which they are parts are represented by distinct vehicles—the surface is represented by the "2½-d sketch", while the relevant physical object is represented by the "3d sketch". It's not clear exactly where Marr draws the boundary between perception and thought though.

Perhaps on other occasions you don't *infer* the existence of a house (because you don't have the requisite background beliefs), but you do entertain a thought or mental image of a house, where this is due to your association of houses with façades. For instance, perhaps you know that you're on a film set and that the façades in your vicinity are not attached to houses, still you can't help but entertain house-thoughts or house-images due to your strong association of façades with houses.

5.2 Thoughts and the two-vehicles approach

In addition to the perceptual case, the two-vehicles approach can be straightforwardly applied at the level of thoughts. For instance, if in thinking about the apple's surface I seem to be in a state that represents the apple itself then this is because I directly represent both the apple and its surface in thought. In many cases, this will be because I token a single thought that *directly* represents both the apple's surface and the apple itself—e.g. a thought of the form, *THE APPLE I'M LOOKING AT HAS A GREEN SURFACE*. In other cases, it may be that I directly represent both the apple and its surface in thought because my thought about the apple's surface causes me to token a distinct thought about the apple itself, or vice versa.

5.3 Pictures

Finally, consider the case of pictorial representation. In particular, consider the intuition—assuming there is one—that a painting of a façade represents a whole house. Is it necessary to posit an *indirect* representation of the house in order to accommodate this intuition?

To see why it is not necessary, consider the fact that, just like linguistic intentionality, it is widely held that pictorial intentionality is derivative of mental intentionality, but presumably not in precisely the same way. If this widely held view of pictorial representation is true then it is far from clear that we need to go beyond a commitment to *direct* pictorial representations in order to accommodate the intuition that a painting of a façade represents a whole house.

For instance, take the kind of view according to which pictorial representations represent what they do, in part, because of the intentions of their creators. Perhaps the content-conferring intention is, in part, to cause viewers to token certain types of images or concepts, e.g. I intend to cause viewers of my drawing (which resembles a dog) to token dog-images or dog-concepts.[19] If we adopt this kind of theory of pictorial representation is there any need to invoke *indirect* pictorial representations? I don't see how any such need arises. Consider the painting of the façade which, intuitively, represents a whole house. According to the kind of view of pictorial representation under consideration, the painting *directly* represents a whole house, in part, because the creator intended to cause viewers to token house-images or house-concepts. An appeal to *indirect* pictorial representation is just not needed in order to accommodate the conviction that a house—not just a façade—is represented.

[19] For examples of this kind of approach to pictorial representation see Abell (2005, 2009) and Blumson (2009).

I have only given a brief sketch of one kind of approach to pictorial representation, but it shows that we don't necessarily need to invoke indirect representations in order to capture Kriegel's intuitions about what gets represented in pictorial cases. If Kriegel wants to support the principle of indirect representation via an appeal to cases of pictorial representation then the onus is on him to show that the best theory of pictorial representation requires us to invoke *indirect* pictorial representations of complex wholes over and above direct ones.

6 The advantages of eschewing indirect representations

So we don't need to posit indirect representations in order to accommodate Kriegel's intuition that in the kinds of cases he focuses on, the subject represents both part and complex whole—direct representations of both will suffice. But we can go further than this for, as I'll argue now, my approach is in fact superior to Kriegel's for two reasons: (1) it has greater intuitive support, and (2) it has more power when it comes to explaining behavior.

6.1 Intuitions

We've already seen that even the best versions of the principle of indirect representation, (IR_U*) and (IR_D), have counterintuitive consequences that my alternative approach can avoid. For as I pointed out above, as long as the subject's disposition to directly represent the given complex (upon directly perceiving or thinking about the relevant part) is not manifested, we're not inclined to attribute an occurrent representation of the complex to her. For instance, recall the case in which your disposition to directly represent a molecule upon directly representing the electron (in thought) is not manifested due to some distraction. Given your failure to token a direct representation of the molecule, the intuition is that you don't seem to be in a state that occurrently represents it at all. The two-vehicles approach straightforwardly accommodates this intuition—you don't occurrently represent the molecule because you're not caused to token a direct representation of it—whereas neither Kriegel's original principle, (IR), nor any of the revised versions that we've considered can.

6.2 The asymmetric dependency intuition

So my account of what is going on in those cases in which the subject seems to represent the complex whole by representing one of its parts can accommodate intuitions that Kriegel's approach cannot. Are there any intuitions that Kriegel's approach can accommodate that mine cannot? Here's one reason to worry that this might be the case. Recall Kriegel's observation that in many cases it seems natural to say that the subject represents the complex whole "by", or "in virtue of", representing the part (2009a, p. 225). There seems to be a sense in which the kind of dependency referred to here is asymmetrical.

For instance, consider Berkeley's famous example in which you hear a horse-drawn carriage on the street *by*, or *in virtue of*, hearing the clip-clop of the horse's hooves. It seems as though your hearing of the horse-drawn carriage depends on your hearing of the movement of the horse's hooves, but not vice versa. One could also cast the intuition in terms of counterfactuals by saying that it seems as though, in some sense, if you hadn't heard the clip-clop you wouldn't have heard the horse-drawn carriage, but not vice versa.

The fact that your representation of the horse-drawn carriage asymmetrically depends on your representation of the movement of the horse's hooves seems to be straightforwardly accommodated by Kriegel's one-vehicle approach. On Kriegel's approach, you indirectly represent the horse-drawn carriage in virtue of directly representing the movement of the horse's hooves, but you don't represent the movement of the horse's hooves in virtue of directly representing the horse-drawn carriage—according to Kriegel's account, you don't directly represent the horse-drawn carriage at all in the present case.

Can the two-vehicles approach accommodate the intuition that your hearing of the horse-drawn carriage asymmetrically depends on your hearing of the clip-clop? At first glance, one might worry that it cannot, for in claiming that distinct vehicles represent the movement of the horse's hooves and the horse-drawn carriage respectively, the two-vehicles approach foregoes the tight, asymmetric, relationship between these two representations that exists under the one-vehicle approach.[20]

This worry is short-lived once we recall how the proponent of the two-vehicles approach explains what gives rise to the (additional) direct representation of the complex whole in these kinds of cases. As I argued above, the cases in which the subject seems to represent the complex whole by representing the relevant part are those cases in which the subject's direct representation of the part *causes* her to directly represent the complex whole. More specifically, I gave two credible ways in which this might happen: (i) the subject's background beliefs concerning complexes of the relevant kind, along with her representation of the given part, causes her to (unconsciously) infer that the part is integrated into the given complex, or (ii) the subject associates objects of the relevant kind with complexes of the relevant kind and thus her representation of the part causes her to token a direct representation of the complex whole (but not via an inference). Applying this to the present case, we can say that your hearing of the clip-clop *caused* your representation of the horse-drawn carriage. Moreover, given that your representation of the horse-drawn carriage didn't cause your hearing of the clip-clop, the two-vehicles approach can indeed accommodate the intuition that the former asymmetrically depends on the latter.

Furthermore, consider what the counterexamples to (IR), (IR_U*) and (IR_D) given above show about the nature of the asymmetric dependency we're honing in on here. The lesson of these counterexamples was that in the kinds of cases we've been considering, we're only inclined to attribute an occurrent representation of the complex whole to the subject if she manifests her disposition to directly represent it as a result of directly representing the part. This shows that intuitions in fact favor the view that in those cases in which the subject seems to represent the complex

[20] Thanks to an anonymous referee for raising this worry, along with the example from Berkeley.

whole *by*, or *in virtue of*, representing the part, the dependency in question is a *causal* one, and that's a view that is at odds with Kriegel's one-vehicle approach.

6.3 Explaining behavior

So intuitions favor my approach over Kriegel's. But that's not the only reason to favor my alternative. Another reason concerns the explanation of behavior. Take the case of the lemon battery again. Suppose you know that your housemate usually uses your lemons to construct lemon batteries. Moreover, suppose you know that he invariably leaves these lemon batteries on the kitchen floor, and so you're disposed to directly represent a lemon battery upon directly perceiving a lemon on the kitchen floor. Finally let's suppose that before walking past the kitchen and directly perceiving the lemon, you were intending to clean your bedroom immediately. With this in mind, consider the following two scenarios.

In the first scenario, you walk past the kitchen and directly perceive the lemon on the floor but no other parts of the lemon battery (e.g. the wires and electrodes). You token a direct representation of the lemon battery as a result, before stopping and exclaiming, "I hate it when he uses my lemons to make batteries!" Finally, you walk into the kitchen and angrily crush the lemon battery under your foot.

In the second scenario the difference is that when you walk past the kitchen and directly perceive the lemon you fail to manifest your disposition to directly represent the lemon battery for some reason (e.g. a loud noise distracts you). Moreover, instead of getting angry and walking into the kitchen to crush the lemon battery, you continue on to your bedroom and satisfy the intention you had to clean your bedroom immediately.

Now, when it comes to the two scenarios just described, we have the same perceptual state type (i.e. the state type of which the initial perception of the lemon is an instance in each scenario) leading to two different pieces of behavior. The question is, why? Why do you say, "I hate it when he uses my lemons to make batteries!" before angrily crushing the battery in the first scenario, but not in the second scenario wherein you satisfy your intention to clean your bedroom immediately without taking any further interest in the lemon?

The divergence in behavior is easy to explain under the two-vehicles approach. The explanation is that in the first scenario you directly represent the lemon battery in addition to your direct (perceptual) representation of the lemon-part, which leads to your angry behavior, whereas you don't represent (in either perception or thought) any parts of the lemon-battery-complex other than the lemon in the second scenario and so you don't exhibit the same lemon-crushing behavior. Of course, this explanation of why your behavior differs in the two scenarios implies that (IR), (IR_U*) and (IR_D) are all false, for according to these versions of the principle of indirect representation you (occurrently) represent the lemon-battery-complex in *both* of the scenarios described above, not just the first one.

What could Kriegel say in reply to this challenge? The only thing he could say is that you do indeed represent the lemon-battery-complex in both scenarios, but that you don't exhibit any relevant lemon-battery-behavior in the second scenario because, unlike in the first scenario, you don't *directly* represent the lemon-battery-complex.

The obvious problem with this reply is that it is tantamount to conceding that indirect representations are causally and explanatorily impotent when it comes to behavior, and this is certainly a strike against Kriegel's one-vehicle approach, as well as the modified approaches based on (IR_U*) and (IR_D). At the very least, an alternative route to motivating the principle of indirect representation is thereby cut off, for Kriegel can't claim that, intuitions aside, the notion of indirect representation gives us an increase in explanatory power when it comes to behavior.[21,22]

7 Conclusion

According to Kriegel, a conscious state represents itself in virtue of the fact it is a complex state such that one of its parts—the higher-order part—represents the relevant lower-order part and thereby indirectly represents the whole state. In order to motivate the appeal to indirect representation, Kriegel uses examples from perception and pictorial representation in an attempt to support the principle that if *X* represents *Y* and *Y* is highly integrated into *Z*, then *X* thereby indirectly represents *Z*. However, as I have argued, Kriegel's principle of indirect representation is problematic for a number of reasons.

First, I argued that there are cases in which a representation of the complex whole does not seem to go along with a representation of the relevant part—in particular, cases in which the subject doesn't expect the complex to be present. The lesson drawn from these counterexamples was that in attempting to fix Kriegel's principle of indirect representation, facts about the subject's expectations cannot be bracketed when specifying the conditions under which she represents the complex whole by representing one of its parts.

I then considered two strategies for amending (IR) by taking the subject's expectations into account. Under the first strategy, the subject must believe that objects of the relevant kind are usually highly integrated into complexes of the relevant kind; while under the second, the subject must be disposed to represent the complex as a result of directly representing the part. However, as I argued, these versions of the principle of indirect representation face their own counterexamples—cases in which it doesn't seem as though the subject is in a state that occurrently represents the complex whole because her background beliefs concerning the relevant kind of complex lay dormant.

I then suggested alternative strategies for accounting for those cases in which the complex whole is represented in virtue of the fact that the relevant part is directly

[21] As I suggested above in Sect. 4, the fact that indirect representations don't have any role to play in explaining behavior might partly explain why it is that we're disinclined to attribute a representation of the given complex to the subject in those cases in which she fails to directly represent it as a result of directly representing one of its parts.

[22] Even though I've focused on the perceptual case to make the point that positing indirect representations won't help us to explain behavior, I could just as easily have appealed to scenarios involving thoughts or pictures. That is, I could've appealed to a pair of scenarios—analogous to the pair given above—in which you directly represent the lemon in thought as opposed to perception, or a pair of scenarios in which you directly represent the lemon pictorially.

represented, where these alternatives eschew the notion of indirect representation. Moreover, I argued that these alternative approaches enjoy the following advantages over those approaches that invoke indirect representations: (1) they are more faithful to our intuitions about the contents of our perceptions, thoughts and pictures, and (2) the notion of indirect representation gives us no increase in explanatory power when it comes to behavior.

The moral is that Kriegel's theory of consciousness requires an appeal to a notion of indirect representation that lacks independent motivation and has counterintuitive consequences regarding what we represent when we directly perceive, think about or picture parts of complex wholes. Whether these considerations are enough to warrant abandoning the self-representational theory in favor of some other theory of consciousness will depend on one's assessment of how the other arguments for and against the self-representational theory fare. Nonetheless, I think my arguments cast serious doubt on the claim that a naturalistically respectable version of the self-representational theory of consciousness is viable.

Acknowledgments I'm especially grateful to David Rosenthal for very useful comments on earlier versions of this paper. I would also like to thank an anonymous referee, Jacob Berger, Ryan DeChant, Gary Ostertag, Jesse Prinz and Elmar Geir Unnsteinsson for helpful comments on earlier drafts. A version of this paper was presented in the spring of 2012 as part of the *Brown Bag Lunch Talk Series* hosted by The Saul Kripke Center at The Graduate Center, CUNY. I'd like to thank the audience for that talk for useful feedback.

References

Abell, C. (2005). Pictorial implicature. *The Journal of Aesthetics and Art Criticism, 63*(1), 55–66.

Abell, C. (2009). Canny resemblance. *Philosophical Review, 118*(2), 183–223.

Armstrong, D. M. (1968). *A materialist theory of the mind*. London: Routledge.

Armstrong, D. M. (1984). Consciousness and causality. In D. Armstrong & N. Malcolm (Eds.), *Consciousness and causality*. Oxford: Blackwell.

Blumson, B. (2009). Defining depiction. *British Journal of Aesthetics, 49*(2), 143–157.

Brogaard, B. (2012). Are conscious states conscious in virtue of representing themselves? *Philosophical Studies, 159*(3), 467–474.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Carruthers, P. (1996). *Language, thought and consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge: Cambridge University Press.

Carruthers, P. (2005). *Consciousness: Essays from a higher-order perspective*. Oxford: Oxford University Press.

Gennaro, R. (2002). Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy, 32*, 293–330.

Gennaro, R. (2004). Higher-order thoughts, animal consciousness, and misrepresentation: A reply to Carruthers and Levine. In R. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 45–66). John Benjamins.

Gennaro, R. (2005). The HOT theory of consciousness: Between a rock and a hard place? *Journal of Consciousness Studies, 12*(2), 3–21.

Gennaro, R. (2006). Between pure self-referentialism and the (extrinsic) HOT theory of consciousness. In U. Kriegel & K. W. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 221–248). Cambridge, MA: MIT Press.

Gertler, B. (2012). Conscious states as objects of awareness. *Philosophical Studies, 159*(3), 447–455.

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. W. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 143–170). Cambridge, MA: MIT Press.

Kriegel, U. (2007). A cross-order integration hypothesis for the neural correlate of consciousness. *Consciousness and Cognition, 16*, 897–912.

Kriegel, U. (2009a). *Subjective consciousness: A self-representational theory*. New York: Oxford University Press.

Kriegel, U. (2009b). Self-representation and phenomenology. *Philosophical Studies, 143*, 357–381.

Kriegel, U. (2012a). In defense of self-representationalism: Reply to critics. *Philosophical Studies, 159*(3), 475–484.

Kriegel, U. (2012b). Self-representation and the explanatory gap. In J. Liu & J. Perry (Eds.), *Consciousness and the self: New essays* (pp. 51–75). Cambridge: Cambridge University Press.

Levine, J. (2010). Review of Uriah Kriegel, subjective consciousness: A self-representational theory. *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/24315-subjective-consciousness-a-self-representational-theory/

Lycan, W. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359. Reprinted in *Consciousness and Mind,* 21–45.

Rosenthal, D. (1990). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: philosophical debates* (pp. 729–753). Cambridge, MA: MIT Press.

Rosenthal, D. (1993). Thinking that one thinks. In M. Davies & G. W. Humphries (Eds.), *Consciousness: Psychological and philosophical essays* (pp. 197–223). Oxford: Basil Blackwell. Reprinted in *Consciousness and Mind* (pp. 46–70).

Rosenthal, D. (2002). Explaining consciousness. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary* (pp. 406–421). New York: Oxford University Press.

Rosenthal, D. (2005). *Consciousness and mind*. Oxford: Oxford University Press.

Rosenthal, D. (2011). Exaggerated reports: Reply to block. *Analysis, 71*(3), 431–437.

Spelke, E. (1988). Where perceiving ends and thinking begins: The apprehension of objects in infancy. In A. Yonas (Ed.), *Perceptual development in infancy* (Vol. 20)., The Minnesota Symposium on Child Psychology Hillsdale, NJ: Lawrence Erlbaum.

Van Gulick, R. (2012). Subjective consciousness and self-representation. *Philosophical Studies, 159*(3), 457–465.

Weisberg, J. (2008). Same old, same old: The same-order representational theory of consciousness and the division of phenomenal labor. *Synthese, 160*, 161–181.